In presenting the dissertation as a partial fulfillment of the requirements for an advanced degree from the Georgia Institute of Technology, I agree that the Library of the Institute shall make it available for inspection and circulation in accordance with its regulations governing materials of this type. I agree that permission to copy from, or to publish from, this dissertation may be granted by the professor under whose direction it was written, or, in his absence, by the Dean of the Graduate Division when such copying or publication is solely for scholarly purposes and does not involve potential financial gain. It is understood that any copying from, or publication of, this dissertation which involves potential financial gain will not be allowed without written permission.

7/25/68

THE PSYCHOMETRIC EVALUATION OF AN INSTRUMENT TO ASSESS COLLEGE TEACHING CLASSROOM EFFECTIVENESS

A THESIS

Presented to

The Faculty of the Division of Graduate

Studies and Research

by

Albert Perry Schwartz

In Partial Fulfillment

of the Requirements for the Degree

Master of Science in Applied Psychology

Georgia Institute of Technology

November, 1972

THE PSYCHOMETRIC EVALUATION OF AN INSTRUMENT

TO ASSESS COLLEGE TEACHING CLASSROOM EFFECTIVENESS

Approved:

Chairman

Date approved by Chairman: 10 November 1972

ACKNOWLEDGMENTS

I am very grateful to my thesis committee, Dr. William Ronan, Dr. James Walker, and Dr. Dale Baskett, for their invaluable assistance to me during the period of my research. Without their help this thesis could not have been completed. I also extend thanks to Dean Sam Webb, whose advice contributed substantially to the research design and execution. Finally, to those students and professors who participated in the research goes my deepest appreciation.

This thesis is dedicated to Valerie.

TABLE OF CONTENTS

TABLE OF CONTENTS (continued)

LIST OF TABLES

LIST OF TABLES (continued)

CHAPTER I

INTRODUCTION AND LITERATURE REVIEW

A report by Kenneth Eble (undated), the director of the Project to Improve College Teaching which was created jointly by the American Association of University Professors and the American Association of Colleges, summarized many of the issues arising from the need to assess college faculty:

> Evaluation is a loaded word. Faculty members are no different from other human beings if they stiffen slightly at the prospect of being evaluated. How does one go about evaluating as complex and personal an act as teaching anyway? If we could evaluate teaching, would it lead to improvement? And how can we either evaluate or improve teaching if we don't know what good teaching is? (p. 8)

This research explores, substantially, the questions raised by Dr. Eble.

Colleges and universities have faced for years the need to evaluate teacher effectiveness of their faculty, and, indeed, this evaluation does occur. It is evident, however, that the vast majority of institutions of higher education in the United States have not sought or employed objective methods of performing such assessments. Gustad (1961) determined that at almost every one of 584 colleges he surveyed, hearsay was the source of information used in judging teacher performance; McGrath (1968) similarly concluded that college faculty promotions were

based most heavily on the personal idiosyncracies of those with the authority to promote despite the fact that promotion was presumably a reward for good teaching. Over the period from 1961 to 1964, the administrative use of methods which gather information systematically from students about the classroom performance of the faculty declined at the institutions which Gustad (1964) surveyed.

In a comprehensive review of the literature, Ronan (1971) presented evidence that almost every attempt at systematizing teacher evaluation has employed ratings as teacher performance criteria. Research by Domas and Tiedeman (1950), Barr and Jones (1958), and Eells (1967) confirmed this view. There is no indication that the use of rating methods in teacher evaluation research is declining.

Remmers (1963) discussed this massive reliance on ratings as teacher performance criteria:

> The reason for these conditions is readily understood. Many of the variables in research on teaching are so complex that tests, questions, and objective behavior records are either inadequate or too inconvenient. (p. 329)

Remmers then presented the various types of rating scales and the properties of these scales. His review points to the conclusion that ratings have very little to recommend them except their expedience. In fact, Remmers' summary noted that research on raters and ratings is needed before they can be accepted as performance criteria data sources.

Ratings as Performance Criteria

The most common criterion of individual performance used in personnel research studies has been and continues to be a rating by a designated observer of the individual performance. A rating is defined as a question or group of questions which demand that the observer make a subjective judgment about a ratee's position on a continuum which is at best vaguely defined. Performance quality is usually recorded on some sort of scale--graphic, forced-choice checklist, or on some refinement of these. An example of a rating is, "This professor is better than the average. Yes or No?" The observer must make judgments about the meaning of the word "better" and the identity of the "average" professor. In contrast, "This professor is on time for every class. Yes or No?" asks no subjective judgment from the observer and is not a rating.

Cureton (1951) noted that such distinctions between subjective and objective observations are imperfect. He commented:

> In the final analysis, of course, all observations are subjective. But operations such as counting, adding, recording the times at which a given series of acts by the person observed began and ended, and the like, possess such a high degree of objectivity that they are commonly termed "objective." Even the observations of Alice's acts of kindness and unkindness are somewhat "objective," though a certain minimum of interpretation (and hence of subjectivity as we are using the term here) is necessary. We shall term an observation "objective" whenever there is a generally agreed-upon standard which can be communicated from one observer to another with little or no ambiguity, and when the observer is not required to interpret or judge, but only to record his observations. (p. 637)

Considerable research attention has been devoted to developing the most perfect rating form or method, yet relatively little notice has been given to ascertaining exactly what data are actually obtained by performance ratings. The following sections are concerned with determining whether ratings of performance yield acceptable judgment of performance excellence.

The general performance criterion problem pervades psychology. Several clinical psychologists have given special attention to the topic. Goldberg (1967) summarized the literature indicating the inadequacies of clinical diagnoses, and both Mischel (1968) and Krasner (1971) commented on the inaccurate nature of clinical assessments. In particular, they noted that there exists a general failure to obtain exact objective behavioral specifications for both diagnoses and prognoses.

The literature cited in the following sections was taken largely from the fields of education and industry and was limited to two types of studies. The first type are those studies in which criteria were established by independent judges or groups of judges who rated the same performance. The second type are studies in which both ratings and objective or bookkeeping criteria of the same performance were gathered. Bookkeeping performance criteria are those which are gathered by a counting process such as words-per-minute typed by a secretary. Focus was restricted to these two types of studies in order to determine the adequacy of observers' ratings of performance as indicators of per-

formance excellence.

Independent Ratings of the Same Performance

If a performance measuring device is to yield scientifically acceptable results, an indispensable condition is that the results be replicable. With performance ratings, this demands that independent observers of the same performance must agree in their assessments of the performance.

The following paradigm would be appropriate for determining whether such agreement exists among independent observers. Two or more people or groups of people should rate the same performances of people engaged in some activity. Furthermore, these ratings must be undertaken so that the assessments reported by one observer or group of observers are in no way dependent upon the assessments made by other observers or groups of observers. For instance, separate groups of observers should not have any opportunity to discuss the performance to be rated before the ratings occur.

In the practical situation the requisite condition of complete rater independence is difficult to obtain. For example, numerous studies employing ratings were reported from the various military services where noncommissioned officers and the immediate commanding officer rated the same persons' performances. The ratings by the noncommissioned officers very possibly were not independent of those by the commanding officer because the raters probably discussed the ratees at various

times in the course of normal duty. To some extent any rating was a product of these discussions. This same lack of independence of the various groups of raters is present in many other types of multi-rater studies; in general, overall multi-rater studies, it is usual to find inter-rater correlations on the order of 0.60, but, as will be seen in the following literature, correlations are generally much lower when raters are making assessments independently of each other.

The rest of this section reports research in which independent raters or groups of raters evaluated the same performance. It must be noted that it was often difficult to determine the degree of rater independence in the reported research. In some cases it was obvious that independence existed among raters, but for most cases complete independence probably did not exist. However, there was a degree of rater independence that was thought to be very close to total in all studies reported hereafter.

The first of the following studies is from the field of education. Evaluation of teacher performance by rating has quite a long history; dozens of studies have been reported (Ronan, 1971). What is presented is merely a sample of all such studies.

In an early study, Heilman and Armentout (1936) used the Purdue Rating Scale for Instructors to obtain performance evaluations on 46 different teachers in fifty different classrooms at several institutions; a total of 2,115 ratings were gathered. Wide differences in the results

among students rating the same teacher were found; standard deviations of such ratings varied from 4.25 to 18.45 on the 100-point scale employed in this research. Teachers were also rated quite differently by different classes; standard deviations of these ratings ranged from 4.75 to 27.30 on the same 100-point scale. The data were so inconsistent both within and over classes for any given teacher as to make the authors conclude that evaluation of that teacher was impossible. Thus, whether viewing each student individually as an independent rater of teacher performance or viewing each class as an independent group of raters of teacher performance, the conclusion must be drawn that independent observers could not agree on the appropriate value of the ratings.

A study reporting ratings of teacher effectiveness in which raters were probably as independent as it is possible for them to be was that of Cook and Leeds (1947). Teacher performance was evaluated by ratings from principals, teaching experts such as education professors, and students. The rating intercorrelations were: (1) principals versus experts, 0.48; (2) principals versus students, 0.39; and (3) experts versus students, 0.33. It is obvious that there was very little agreement among the three different groups of raters about the performance of the rated teachers.

Crawford and Bradshaw (1968), in a study with implications similar to Cook and Leeds, had some 300 students write themes describing effective classroom teaching; and from these themes, 13 categories of

such statements were derived. For example, the most frequently reported category was, "Has thorough knowledge of subject matter plus substantial knowledge in related fields." These categories were paired so that all unique combinations of two at a time were constructed, and these pairs were made into a checklist. The one item of each pair considered more essential for effective teaching was then chosen by 50 faculty members, 158 students in psychology, and 30 school administrators. It was found that judges within any given group agreed substantially on their rankings. Greatest consistency of opinion was found within the administrative group; lowest consistency was found within the group of female students. There were wide differences among how the groups ranked the various items; chi-square tests of independence indicated differences among groups were generally significant ($p = 0.05$).

The final study from the education literature which bears on the findings of Cook and Leeds was by Webb (1967). For this research, both students and teachers rated 13 course goals such as teaching of knowledge with regard to the following three continuums: (1) importance of accomplishing each goal; (2) relevance of each goal; and (3) emphasis given each goal. The major purpose was to estimate the correspondence between students and faculty in their ratings. Fourteen classes totaling 324 students participated in the study; teachers rated themselves on each goal, and students rated the teachers. The findings implied that there is considerable lack of congruence between what instructors say

they are attempting and students' perceptions of these attempts.

These studies from the field of education all indicate that independent raters of teacher performance disagree in their observations. In fact, the research indicates that the various independent groups associated with the school setting must disagree in how they rate teacher performance because they have different views of what the teacher does and should do. When rating, each group rates with regard to a different anchor.

The rest of the studies presented in this section are from the field of industrial psychology. In industrial psychology there have been many studies in which ratings have been completed by independent groups of observers of the same performance. The following studies are a sample of these.

One early study investigating the general question of the relationship between ratings by independent groups was that of Springer (1953). She compared performance ratings of 100 candidates for promotion made by supervisors with those made by coworkers. A graphic, 8-item scale was used by 68 foremen and assistant foremen for the ratings. Peer ratings correlated with superior ratings in the range of 0.15 to 0.39. Obviously, the two groups of ratings could not be considered equivalent.

In research with intent similar to that of Springer, Besco and Lawshe (1959) had foremen rated on structure and consideration by su-

pervisors and by subordinates. Superior evaluations showed negligible correlations with subordinate perceptions of the same attributes ranging from -0.15 to 0.08. It was concluded that superior and subordinates view leadership behaviors in different and unrelated ways.

In another study comparing superior with subordinate ratings, Strander (1965) had managers rated twice in a three-year period. Both subordinate and superior ratings were taken each time; for the most part, managers were rated by the same people in both instances. A total of 19 different variables were assessed and factor analyzed. Seven major factors and several specific factors were extracted. Some illustrative findings were: (1) subordinates considered manager mental alertness very important and supervisors did not; (2) planning and organizational ability received very high loadings from supervisors but virtually zero loadings from subordinates; (3) human relations skill was considered important by supervisors but not by subordinates; and (4) ability to promote group cohesiveness showed high agreement among subordinates but no agreement among supervisors. In general, there was little agreement between supervisors and subordinates on either the importance of the variables or who possessed them.

These same results are constantly repeated in the literature. Superiors, performers themselves, peers, and subordinates do not agree on ratings of performances. Further illustration of this lack of agreement is evidenced in the studies by Kirchner (1966), who found self-

ratings only marginally correlated with superior ratings (0.08 to 0.33); Charest, Cowart, and Goodman (1969), who found little agreement among superior, self- and peer ratings; Kavanagh, MacKinney, and Wolin (1971), who found superiors and subordinates disagreed on ratings of managers; and Schneider and Bartlett (1970), who found superiors and peers ratings were mostly inconsistent with each other. Many more such studies exist in the literature.

Prien (1962) and Prien and Liske (1961) have attempted to account for this disagreement among independent raters or groups of raters in the industrial situation. They found that superiors do not agree with peers of the people being rated about just what the given job actually entails. That is, they are actually rating two different jobs when they rate performance on what is supposedly a unitary, well-defined job. These results are consistent with the results of the study by Webb (1967) which was previously cited in the discussion of the education literature.

An elegant experimental study by Lifson (1953) serves very well to clarify the research presented in this section. In the study, trained time-study personnel rated work pace on a continuum from slow to normal to fast. Each rater judged the pace of five different people on four different jobs. Each of the workers was rated twice at one-month intervals. The workers were actually students who performed after considerable practice paced by a metronome. The study revealed several

sources of considerable error in ratings. One-third of the variance was accounted for by inter-rater differences. Some jobs were rated more reliably than others; some workers were consistently rated more reliably than others. Statistically significant ($p = 0.05$) interactions among raters, workers, and jobs accounted for a considerable portion of the variance. In effect, the results indicated that even with persons specifically trained to rate the work performance of others, they could not agree among themselves about what they were actually observing with people at work.

In concluding this section, the research concerned with independent raters evaluating the same performance indicates that such raters show very little agreement in their assessments of performance. This is found with a variety of raters, tasks, and evaluation instruments. It should be reemphasized here that these results are from research where, at least relatively speaking, the ratings were made independently. As previously indicated, rater independence is often difficult to judge from the research reports, but where it is shown they are independent, the rater-to-rater correlation rarely exceeds 0.20. This, in fact, would probably represent the approximate median value if all such studies were to be tabulated and, of course, represents a small portion of the performance variance.

Ratings and Objective Measures of Performance

If independent performance ratings are mutually incongruent as

was shown in the previous section, it is still possible that ratings show convergent validity (Campbell and Fiske, 1959). In this case, convergent validity would be evidenced when performance ratings are consistent with more objective measures of the same performance. For many purposes, it would be sufficient for ratings to have convergent validity only. This is especially the case when it is too inconvenient or too impractical to gather objective performance indices.

The research to be presented was taken from studies in which both ratings and one or more objective indices of performance excellence were used as performance criteria. These studies indicated that relationship between the subjective indices of performance excellence, that is, ratings, and the objective indices of performance excellence is weak at best. Two studies from the field of education are presented first; these are followed by several studies from the industrial psychology literature.

The first of the education studies is by Elliot (1950) who used an amended version of the Purdue Rating Scale for Instructors, adding two items: one soliciting comparison of a particular instructor with others the student had encountered, and the other asking that, if the occasion arose, should the particular instructor be replaced. Subjects of the study were instructors in the Chemistry Department at Purdue University. A teaching attitude test, How to Teach and Learn in College, was also administered to the instructors. The objective performance mea-

sure for each instructor was the average discrepancy between the actual grades and predicted grades of the students in his course. The predicted grades were based on a multiple regression equation from the American Council on Education Psychological Examination. Actual grades were based largely on objectively scored tests. The results showed that only 5 of 24 possible correlations were statistically significant ($p = 0.05$) and all correlations were quite low. Neither instructors' knowledge of the subject matter as measured by an achievement test nor instructors' scores on the attitude scale were related to the objective performance criterion. Rated teacher effectiveness was essentially incongruent with the objective teacher performance criterion.

The second education study was conducted in a military setting; Borg and Hamilton (1956) performed the study at an Air Force basic training school. On several different traits, 89 instructors were rated by students, were self-rated, were rated by superiors, and were rated by peers. The objective performance criterion for a given instructor consisted of his students' grades on twelve standardized problems. The correlations of the ratings with the objective criterion were the following: (1) student ratings, 0.19; (2) peer ratings, 0.11; (3) supervisors, 0.13; and (4) self, 0.01. None of the ratings were significantly ($p = 0.05$) related to the objective performance criterion.

There have been other studies in the education literature similar to those mentioned, but few so extensive and meticulous. The same

general findings recur in virtually all of them: Ratings show little or no relation to objective measures of performance.

The studies that follow were done at industrial settings. Basically, they confirm the results of the studies cited previously.

One of the earliest of the industrial studies bearing on the relation between subjective and objective indices of performance, Braunhausen (1929), correlated supervisors' ratings of their subordinates' job performance against the subordinates' grades on a job sample test administered to them. For the two different groups used in the study, Yule's coefficients of association were 0.14 and 0.56. That is, the ratings accounted for less than a third of the variance present in the job sample test. More recently, Gaylord, et al. (1951), found that each of three objective indices of performance by file clerks correlated less than 0.56 with supervisors' ratings of the clerks' performance.

A striking example of the lack of relationship between job merit ratings and objective indices of job performance was evidenced in a study by Tiffin and McCormick (1965). In this study a test was administered which consisted of having the inspectors find defects in a sample with a known number of the different types of possible defects. The inspectors were then rated by their superiors. Results were the following: Inspectors rated good scored 77 per cent correct on the test; inspectors rated average scored 79 per cent correct; inspectors rated fair scored 84 per cent correct; and inspectors rated poor scored 68

per cent correct. Thus, the most highly rated inspectors were not the best as measured by the job sample test.

There are many more studies in the industrial literature with results similar to those previously cited in this section. A small sample of these include the following: Siegel (1954) found rho values between superiors' ratings of Navy craftsmen and job sample checklists used to grade these craftsmen's work to range from 0.14 to 0.66; Fleishman, Harris, and Burtt (1955) found proficiency ratings of factory workers correlated with absences at -0.07, with accidents at 0.13, with grievances at 0.13, and with turnover at 0.24; Prien (1969) found only 3 of 48 possible correlations between ratings and objective indices of performance for bank tellers to be statistically significant ($p = 0.05$); and Baehr and Williams (1968) found sales managers' ratings of their salesmen's performance to correlate with ten-year sales volume at 0.37, with highest year sales volume at 0.37, with route difficulty at -0.20, and with tenure at -0.13.

In concluding this section, it should be noted that the relation between ratings and objective indices of performance excellence has been repeatedly investigated. These investigations employed a great variety of objective performance criteria, but in all cases correlations between the ratings and the objective indices were quite low, rarely exceeding 0.40. Where the 0.40 level was exceeded, it was often the case that raters had access to the objective scores before they were called upon

to rate. It appears safe to infer that ratings do not possess convergent validity; they do not give a true picture of actual performance.

Factor Analyses of Performance Criteria

This section reviews some research which employed factor analyses of intercorrelation matrices of variables which represented ratings and objective measures of performance. It usually happens that subjective measures of a given performance load most heavily on one set of factors while objective measures of the same performance load most heavily on factors orthogonal to the first set. That is, objective performance measures are largely accounted for by underlying variables independent of those which largely account for the subjective performance measures.

The first study, from the education literature, was by Locke (1963) and was concerned with the performances of 122 high school students in a special program at Cornell University. The measures of the students' performances in the program were the following: (1) ratings by their high school teachers; (2) ratings by their instructors at Cornell; (3) self-ratings; (4) a systematic interview of their high school teachers; (5) an objective measure of their productivity in the program; and (6) a measure of productivity quality. A factor analysis yielded two factors. One had high loadings, that is, larger than 0.35, by grades and by ratings given by instructors at Cornell; the other factor had high loadings by productivity, by quality, and by self-ratings. In effect, ac-

tual performance in the program was independent of instructor evaluation of that performance.

The next studies are from the industrial psychology literature. Ronan (1963a) studied performance of skilled tradesmen using three ratings and eight objective measures of performance. Four factors were extracted from the eleven measures. One of these was a promotion-supervisory rating factor; another was an apprentice school rating-grade factor. All of the objective performance factors were independent of these. In a similar study Ronan (1963b) evaluated all workers in two plants with seven objective performance measures and with one rating. Again four factors were extracted from the data at each of the two plants. In both cases loadings for the supervisory ratings were somewhat broadly distributed across the factors but were related largely to absence and disciplinary actions in the one plant and to lost time accidents in the other. That is, many of the objective performance indices were accounted for independently of the ratings of those same performances.

There exist many more factor analytic studies in the same vein as those already cited. Generally, they all reinforce the implication that ratings, subjective measures of a performance, are accounted for by different underlying variables from objective measures of the same performance. This may be interpreted as further evidence that ratings lack convergent validity and that ratings are unacceptable as perfor-

mance criteria.

Difficulties Involved in Observing Performance

Kipnis (1960) and Taft (1955) have discussed some of the major difficulties that are involved in the observation of human performance. Although Kipnis refers to ratings of performance excellence and Taft refers to the ability to judge others, both seriously question the usefulness of human observations of the performance of others. Taft emphasized that there exist traits within each observer, such as intelligence or ignorance, emotional stability or instability, social skills or its absence, and self-insight or the lack of it, which affect the observer's perceptions. Kipnis goes further and emphasizes that the observer interacts with the environment at the time judgments are made; this interaction acts to distort perception. Jones and deCharms (1957) and Jones and Thibaut (1958) have also noted that such interactions definitely take place.

The studies by Rosenthal (1966) have shown dramatically that the observation of a given person is indeed the product of his environment as well as his own perceptions. These experiments have demonstrated that observers generally perceive what they expect to perceive and that such perceptions are easily altered by changes in the environment. Because ratings usually require more complex judgments than those made by the participants in Rosenthal's experiments, it is evident that raters can rarely, if ever, give true readings of performances they are wit-

nessing.

On a logical basis, it seems to be asking raters to do the impossible when requiring them to rate the performance of others. Even the simplest jobs have several dimensions. There are usually several persons rated, and the rater must simultaneously take all of this into account when rating a single individual (Ronan and Prien, 1971). The research evidence seems to show that the perceptual field is so complex that raters cannot in fact do the task successfully (Bruner, 1958). In general, performance ratings lack any kind of demonstrable validity as shown in this research. If their use is to be continued, research effort needs to be devoted to the question of just what raters do rate; it is patent from converging evidence that it is not performance.

Perhaps what is usually rated is made explicit in a quote from Thorndike (1949) in summarizing experience with ratings in the Army Air Force during World War II:

> The trainees were rated by their instructors on approximately ten traits. These same instructors indicated what they considered to be the importance of each of the traits for over-all effectiveness in the job assignment. Though "likableness" was consistently placed at the very bottom of the list in importance, it nevertheless fell at the top in terms of its correlation with an over-all rating. Though the raters disclaimed its importance, it still provided the chief basis for their over-all evaluation. (p. 151)

A more recent study by Graham and Calendo (1969) is related to Thorndike's quotation. These investigators had 69 clerical workers complete Gough's Adjective Check List and Ghiselli's Self-Description

Inventory. From the two instruments, scores for Aggression, Autonomy, Deference, Endurance, Self-Confidence, and Self-Control were derived. Supervisory ratings were obtained from the regular program of evaluation in practice or were specifically completed for new employees. The ratings consisted of five scales for job performance behaviors and eight scales for personal characteristics. Of 40 correlations of ratings versus job performance behaviors, only 2 were significant. However, of 64 correlations of personality trait scores versus ratings of personal characteristics, 30 were significant at the 0.05 level or higher. Graham and Calendo concluded that workers were rated on the basis of personality characteristics, especially conformity, and not on the basis of work performance.

Because it is not known just how raters make their assessments nor indeed what they actually assess, it would seem that ratings should not be used as performance criteria. This research has therefore been designed to assess teacher effectiveness without the use of ratings.

An Alternative to Ratings for Evaluating Teachers

Barr, Bechdolt, Coxe, Gage, Orleans, Remmers, and Ryans (1953) stressed the need for finding observable teacher behaviors which are effective or ineffective with regard to acknowledged goals of the college. Once these behaviors are determined, they can serve as criteria for assessing teachers. A teacher would either exhibit or not exhibit a given behavior; no ratings would be necessary.

Flanagan (1949a, 1949b, 1954) devised the critical incident technique as a direct procedure for recording observations of human behaviors which lead to success or failure with regard to the accomplishment of a specific task such as teaching. In attempts to uncover such behaviors, Smit (1951), Konigsburg (1954), Douglas (1968), Deshpande, Webb, and Marks (1970), and Ronan (1971) employed the critical incident technique soliciting effective and ineffective college teacher behaviors. The lists of behaviors produced in these studies were generally in accord with each other despite variations in populations sampled.

Latham (1969) noted that critical incidents can be readily transformed into statements descriptive of actual job behaviors which may be used to delineate the characteristic behaviors of any individual on that job. The observer is simply required to check whether any given behavior as represented by a statement did or did not occur. No judgment of status relative to some trait, either clearly or poorly defined, is demanded. That is, no performance ratings are solicited. The incidents of teacher behaviors mentioned earlier have been transformed into such statements.

However, the gathering of critical incidents and the compiling of them into a checklist cannot be viewed as the final step in devising a suitable personnel evaluation checklist. The critical incident technique suffers from inherent weaknesses which necessitate the confirmation of any given incident's relation to the accomplishment of the task in ques-

tion. The first weakness is that the degree of the relation is not ascertained merely by gathering critical incidents. A critical incident should bear strongly on success or failure or should be purged from any checklist. Next, the critical incident report may suffer from unreliability; many chance factors can possibly influence the reporting of any incident (Ronan and Prien, 1966). Finally, factors such as selective attention, group membership, and lack of interest may act to distort offered reports of critical behaviors (Safren and Chapanis, 1960). Thus, it is necessary and methodologically sound to corroborate the results of the critical incident studies of teacher performance with a different measure of teacher effectiveness. The present research attempted such a corroboration.

The Critical Incidents

Of the previously mentioned critical incident studies of teacher behaviors, two are considered the most relevant for this study. Both Deshpande, et al. (1970), and Ronan (1971) collected their incidents at the Georgia Institute of Technology, the setting of the present study. However, only Ronan sampled incidents over the entire institution; Deshpande and his associates confined themselves to the School of Mechanical Engineering. Because this study was aimed at schools at the Georgia Institute of Technology other than Mechanical Engineering, the list of critical incidents gathered by Ronan was chosen as most appropriate.

The Performance Criterion

The present research demanded an immediately available, easily accessible teacher performance criterion so as to be completed in a reasonable amount of time. In restricting the criterion in such a manner, it was recognized that possible penultimate and ultimate criteria were disregarded. However, as Cureton (1951) noted, it is virtually impossible to capture the true essence of the college teacher's broad long-range goals within any set of criteria. Thus, more immediate goals and more encounterable criteria were demanded for this type of research. With the establishment of such goals, the students' progress towards these goals could be measured for students of a given teacher; effectiveness could be inferred from this progress.

In a very complete discussion of teacher effectiveness, Barr, et al. (1953), concluded there exist three major areas of teacher responsibilities. These were: (1) to work with students, (2) to work as part of a staff of teachers, and (3) to work with the community. Eckert (1950) also described faculty responsibility as encompassing much more than classroom teaching. However, he noted that although the teacher has staff duties, the one most important duty is to teach in the classroom. Students tend to view this as the only faculty duty. Ronan (1971) found students almost always reported incidents from within the classroom when asked for behaviors of their best and of their worst professor. Thus, in this research, an appropriate criterion of teacher

performance was sought in the classroom.

In the classroom teaching situation, instructional objectives are usually implicitly or explicitly present. An instructional objective specifies the following: (1) the subject matter to be taught, the goals; (2) the process and materials to be used for teaching; and (3) the method for assessing whether the goals have been reached. The most relevant classroom teacher performance criterion in a situation with an encounterable instructional objective can then be conceptualized as the degree to which the teacher facilitated student progress towards the goals as measured by the method specified in the objective (Mager, 1962).

The most common criterion of this type is conceptualized as the degree to which the teacher facilitates students' achievement in his class. Employing this conceptualization, Douglass (1968) operationally defined her criterion as the grade on a standardized final examination. As a control for aptitude, student grade point average was checked to be equal across all classes and was judged to be so.

Aptitude may not actually be controlled by grade point average, however. Reports by Aiken (1963), Webb (1963), Hills (1964), and Astin (1968) indicate that average grades may not reflect high or low ability of students to any major degree, especially when ability is sharply different from one group of students to the next. In effect, there are sources of variance for grades other than student ability level. Thus, if grades of any sort are to be used as a performance

criterion for assessing teacher effectiveness as related to student achievement, it is necessary to control as many sources of performance variability as possible. The controls used in this study are described in a following chapter.

CHAPTER II

STATEMENT OF THE PROBLEM

The objective of this research was to evaluate the critical incident checklist compiled by Ronan (1971) with regard to each item's perceptual stability and with regard to each item's validity for diagnosing college teacher classroom effectiveness. The reliability of professorial behavior from class to class was also evaluated. Conceptual statements describing the various statistics employed are presented in this chapter; operational statements follow in the next chapter.

Perceptual Stability

Perceptual stability was considered the main measure of reliability for a given item; it is a statistic designed to reflect the degree to which students in a class agree on the behavior their professor exhibits. Average perceptual stability of 0.85 or higher was deemed necessary for considering an item acceptable. The rationale for using 0.85 as the cut-off is discussed in a subsequent section.

Validity

The set of null hypotheses that each critical incident, as represented on the checklist, is not related to teacher classroom performance

was evaluated. The criterion of teacher performance, which was mentioned in the previous chapter, was the average grade in a professor's class adjusted for covariance due to student aptitude and past achievement. For the purpose of this research, 81 analyses of covariance were performed with the 5 per cent level of significance considered appropriate. Using this level of significance, the probability of rejecting the null hypothesis if it were true is reduced to 0.05 or less.

Reliability of Professorial Behavior

The set of null hypotheses that a professor's behavior as represented by each item of the checklist is consistent over classes was proposed and tested. Once again the 5 per cent level of significance was deemed appropriate.

CHAPTER III

PROCEDURE AND METHODOLOGY

Subjects

Sample I

The Acting Director of the School of Mathematics consented to the participation of his department in this research for the winter quarter, 1972. Second-quarter freshman calculus, Math 108, constituted the set from which Sample I was derived. This course was taught by twenty professors in as many sections. Any given class participated in the study only with the consent of the teacher; six professors declined to participate in any phase of the study, and one additional professor declined to release the data for the validation phase.

Although there was no explicit course objective for Math 108, a firm concensus existed about the implicit course objective. This implicit objective was the following: (1) to teach the material on integral calculus contained in the text as specified on a departmental course schedule; (2) to use lectures and the blackboard to teach the course; and (3) to measure progress towards the course goal by a departmental final examination. As will be seen in a subsequent

section, the teacher performance criterion used in this research was derived directly from this course objective.

Sample II

This sample consisted of 24 professors' classes in the Schools of Electrical Engineering, of Physics, and of Psychology. No validation analyses were performed on these subjects' data since no common final was given in Sample II. Only reliability of professorial behavior was assessed.

College Teacher Improvement Checklist

The College Teacher Improvement Checklist, which is found in Appendix A, was used to gather the data. This was compiled from the list of critical incidents gathered by Ronan (1971) which was discussed in a previous section. The checklist was administered in one of two ways dependent upon the professor's wishes. The first type of administration was one in which the professor himself distributed the questionnaire and had his students complete it at home. All data for Sample I were collected in this way. The second type of administration consisted of the professor's having the questionnaire distributed and completed in class. Parts of the data for Sample II were collected in this manner.

All students were asked to respond either, "Yes, the professor in my class exhibited this behavior," or "No, he did not."

As requested by several professors, all data were destroyed when

their use was no longer necessary.

The Teacher Effectiveness Criterion

For each class in Sample I, except one, the average grade on the final examination which was common to all classes was collected. The final examination was in the discussion or problem-solving format and was graded by a committee to insure, as much as possible, that it was indeed comparable across classes.

The admissions office then furnished the average scholastic aptitude test scores, verbal (SAT v') and math (SAT m'), and average high school grade (HSA') for the students in each of the 13 classes.

The actual teacher effectiveness criterion was then the average final grade adjusted for covariation attributal to SAT v', SAT m', HSA'. The analytic details of this adjustment are explained subsequently.

Perceptual Stability

The statistic yielding information concerning reliability of report of behavior by students was the coefficient of perceptual stability, a statistic original with this study. This was not meant to be one of the classical coefficients of reliability; furthermore, it was not a correlation coefficient. For each class the coefficient of perceptual stability for a given question was the fraction of the class giving the modal answer for that question.

Considering answers to a given question to have a binomial dis-

tribution, for a class of 16 which was at least the number of questionnaires collected from most classes, the coefficient of perceptual stability must be approximately 0.85 or higher for the hypothesis that the class mode is due to chance to be rejected at the 0.05 level. Therefore, 0.85 was made the cut-off for a question to be considered acceptable with respect to perceptual stability.

Reliability of Professorial Behavior

It was of interest to see how reliable professorial behavior was from class to class. To assess this, each professor teaching two or four classes had half of his classes randomly assigned to Set A. The other half was assigned to Set B. Then for each question the following procedure was used: For each class the modal class judgment for that question was ascertained. The modal judgment was converted to a number by coding "Yes" as +1 and "No" as -1. Then Vector A was constructed by placing the coded modal class judgments for every class in Set A into the vector, professor by professor. Vector B was constructed in the same manner with the constraint that a given dimension in Vector B represent the same professor as that dimension in Vector A. Both vectors were then normalized, and the Pearson Product-Moment coefficient of correlation was calculated. For the behavior represented by the question, this coefficient was the statistic indicating the reliability of that behavior from class to class.

Validity

Each question was used to separate professors into one of two groups--a "Yes" or a "No" group with respect to that question only. Then for each question an analysis of covariance was performance using the criterion and covariates previously discussed. Analyses were considered mutually independent, and no interaction effects were considered.

CHAPTER IV

RESULTS AND RELATED DISCUSSION

Perceptual Stability

A basic objective of this study was to discover whether the checklist of teacher behaviors used in this research (see Appendix A) exhibited acceptable perceptual stability. Perceptual stability was a statistic original with this research. It must be stressed that this checklist was not a rating; that is, it did not ask the student to make a subjective judgment in relation to a vaguely defined continuum. As Lifson (1953) demonstrated, even expert raters show very little agreement among their ratings; teachers are quite aware that ratings give an unreliable assessment of their performance (Ryans, 1954).

Determination of perceptual stability was made on Sample I and Sample II. Out of 81 questions on the checklist, 50 exhibited acceptable perceptual stability in both samples, that is, exhibited mean agreement of 0.85 or higher. The results for each question are given in Table 1.

Especially in view of Lifson's (1953) results indicating ratings have little perceptual stability, it was remarkable that this behavior checklist exhibited such high perceptual stability. Konigsburg's (1954) results were similar; the teacher behavior checklist used in his re-

Table 1. Within Class Reliabilities

Behavior	Sample I		Sample II	
	Mean Agreement	Standard Deviation	Mean Agreement	Standard Deviation
1. Know Students' Names	.863*	.159	.763	.141
2. Talk with Students outside of Class	.969*	.059	.946*	.077
3. Hold Social Events for Students	.993*	.017	.968*	.049
4. Refuse Students Advice on Personal Problems	.919*	.083	.940*	.062
5. Discuss Extra-Class Issues	.679	.172	.750	.154
6. Encourage (Answer) All Questions	.821	.104	.892*	.110
7. Treat All Students Equally	.935*	.105	.985*	.026
8. Ridicule Students	.862*	.111	.962*	.052
9. Give Individual Help with Course	.976*	.034	.923*	.084
10. Lose Emotional Control in Class	.917*	.091	.976*	.040
11. Harass Students during Quizzes	.976*	.034	.984*	.029
12. Make Threats to Students	.964*	.060	.965*	.064
13. Accept Excuses for Missing Quiz	.880*	.115	.888*	.119
14. Refuse to Recognize Other Viewpoints	.897*	.117	.963*	.043
15. Indicate that Students Are Inferior	.900*	.107	.943*	.099
16. Miss More than Two Classes	.975*	.040	.974*	.080
17. On Time for All Classes	.946*	.084	.938*	.081
18. Arrange Deadlines for Student Convenience	.839	.163	.869*	.143
19. End Class on Time	.823	.142	.925*	.105
20. Distribute a Course Plan	.755	.153	.863*	.135

Table 1 (continued)

21.	Follow Course Plan	.792	.155	.803	.129
22.	Give Examples of Quiz Items	.806	.131	.754	.162
23.	Require and Grade Homework	.876*	.155	.912*	.102
24.	Return Papers Promptly Graded	.889*	.187	.934*	.107
25.	Permit Classroom Disturbances	.765	.150	.751	.110
26.	Make False Statements about Course Requirements	.974*	.048	.979*	.035
27.	Give Excessive Work	.907*	.103	.927*	.076
28.	Ask Student Preference as to Topics Covered	.968*	.044	.869*	.132
29.	Ask for Student Suggestions on Teaching Methods	.831	.129	.743	.168
30.	Encourage Discussion	.736	.163	.763	.140
31.	Appear Well-groomed	.955*	.073	.974*	.151
32.	Speak Distinctly	.906*	.145	.920*	.124
33.	Use Humor	.813	.160	.756	.155
34.	Read Lectures	.903*	.121	.771	.167
35.	Appear Nervous	.944*	.061	.922*	.123
36.	Talk Too Rapidly	.765	.145	.804	.157
37.	Give Disorganized Lectures	.901*	.064	.879*	.106
38.	Look at Students While Lecturing	.939*	.093	.914*	.108
39.	Use Language Students Understand	.910*	.117	.932*	.072
40.	Use Profane Language	.988*	.248	.972*	.048
41.	Stress Important Points	.912*	.078	.849	.151
42.	Use Examples for Illustration	.707	.149	.788	.178
43.	Show Usefulness of Material	.718	.151	.841	.142
44.	Admit Not Knowing Answers	.811	.164	.866*	.128
45.	Use Outside References	.690	.172	.801	.152

Table 1 (continued)

46.	Distribute Hand-outs	.848	.158	.834	.162
47.	Use Visual Aids	.991*	.024	.907*	.090
48.	Know Subject Matter	.922*	.135	.914*	.120
49.	Lecture from Text	.745	.155	.776	.143
50.	Cover All Course Requirements	.911*	.097	.897*	.104
51.	Avoid Trivial Detail	.710	.122	.781	.117
52.	Answer Questions and Work Problems on Request	.983*	.033	.934*	.092
53.	Lecture over Students' Ability	.809	.165	.832	.169
54.	Give False Information about Course	.951*	.044	.968*	.045
55.	Refuse to Explain Course Material	.962*	.051	.949*	.077
56.	Follow Course Schedule	.872*	.111	.872*	.108
57.	Prepare for Class	.964*	.063	.963*	.060
58.	Base Tests on Relevant Material	.862*	.151	.919*	.112
59.	Base Tests on Emphasized Material	.840	.130	.855*	.048
60.	Make Tests Too Difficult	.816	.125	.766	.137
61.	Schedule Quizzes at Regular Intervals	.787	.155	.878*	.124
62.	Make Tests Too Long	.809	.174	.766	.148
63.	Comment on Returned Papers	.922*	.089	.863*	.136
64.	Excuse High Average Students from Final	.980*	.035	.910*	.135
65.	Permit Extra-Credit Work	.877*	.122	.827	.162
66.	Disregard Lowest Test Score	.856*	.138	.965*	.047
67.	Refuse to Explain Grading System	.892*	.124	.924*	.091
68.	Tell How Students Are Graded	.807	.132	.870*	.120
69.	Curve Grades	.765	.174	.810	.145
70.	Return All Papers	.966*	.059	.950*	.104

Table 1 (continued)

71.	Grade All Assignments	.880*	.108	.960*	.062
72.	Give Make-Up Tests at Students' Convenience	.787	.118	.741	.140
73.	Grade on Major, Sex, Athlete, etc.	.988*	.030	.995*	.014
74.	Grade on Class Attendance	.855*	.155	.928*	.123
75.	Grade on Final Exam Only	.962*	.084	.994*	.016
76.	Pass/Fail a Predetermined Percentage	.940*	.078	.955*	.075
77.	Willing to Discuss Grades	.922*	.076	.937*	.083
78.	Derogate Teaching	.970*	.053	.967*	.045
79.	Derogate the Course	.924*	.126	.963*	.048
80.	Indicate Preference for Consulting or Research	.973*	.047	.976*	.045
81.	Criticize Fellow Teachers	.983*	.027	.957*	.076

Sample I: $\underline{n} = 14$.

Sample II: $\underline{n} = 24$.

*Acceptable, above 0.85.

search exhibited average test-retest reliability of 0.72. Thus, it seems that behavior checklists are an improvement over ratings with regard to reliability. It must be noted, however, that further research is necessary to determine whether the present checklist indeed has greater perceptual stability than ratings of teacher performance.

Reliability of Professorial Behavior

One issue which received very little attention in the literature on teacher evaluation was that of the reliability of professorial behavior from class to class. If a professorial behavior is reliable from class to class, an overall assessment of that behavior and its contribution to teacher effectiveness is likely to be possible. However, if professors alter their behaviors from class to class, it seems doubtful that any simple, unitary assessment of teacher behaviors and performance effectiveness is possible; perhaps only assessments made class by class would be appropriate.

Those professors from Sample I and Sample II teaching two or four classes were used to determine coefficients of reliability for professorial behaviors. Table 2 gives the Pearson Product-Moment correlations, the reliability coefficients, for each behavior as represented by the questions on the checklist. Of the behaviors, 54 of 81 proved to have coefficients significantly different from zero ($p = 0.05$). However, only one coefficient was above 0.52. So, although most behaviors

Table 2. Reliabilities of Professorial Behaviors

	Behavior	r
1.	Know Students' Names	.382*
2.	Talk with Students outside of Class	.363*
3.	Hold Social Events for Students	.284*
4.	Refuse Students Advice on Personal Problems	.363*
5.	Discuss Extra-Class Issues	.294
6.	Encourage (Answer) All Questions	.363*
7.	Treat All Students Equally	.262
8.	Ridicule Students	.004
9.	Give Individual Help with Course	.199
10.	Lose Emotional Control in Class	.467**
11.	Harass Students during Quizzes	.352*
12.	Make Threats to Students	.301
13.	Accept Excuses for Missing Quiz	.497**
14.	Refuse to Recognize Other Viewpoints	.352*
15.	Indicate that Students Are Inferior	.321*
16.	Miss More than Two Classes	-.045
17.	On Time for All Classes	.352*
18.	Arrange Deadlines for Student Convenience	.301
19.	End Class on Time	.363*
20.	Distribute a Course Plan	.352*
21.	Follow Course Plan	.359*
22.	Give Examples of Quiz Items	.363*
23.	Require and Grade Homework	.363*
24.	Return Papers Promptly Graded	.506**
25.	Permit Classroom Disturbances	.363*
26.	Make False Statements about Course Requirements	.359*
27.	Give Excessive Work	.398*
28.	Ask Student Preference as to Topics Covered	.301
29.	Ask for Student Suggestions on Teaching Methods	.058
30.	Encourage Discussion	.352*

Table 2 (continued)

31.	Appear Well-groomed	.359*
32.	Speak Distinctly	.430**
33.	Use Humor	.352*
34.	Read Lectures	.294
35.	Appear Nervous	.363*
36.	Talk Too Rapidly	.352*
37.	Give Disorganized Lectures	.293
38.	Look at Students While Lecturing	.363*
39.	Use Language Students Understand	.363*
40.	Use Profane Language	.512**
41.	Stress Important Points	.401*
42.	Use Examples for Illustration	.363*
43.	Show Usefulness of Material	.310
44.	Admit Not Knowing Answers	.210
45.	Use Outside References	.359*
46.	Distribute Hand-outs	.258
47.	Use Visual Aids	.191
48.	Know Subject Matter	.273
49.	Lecture from Text	.359*
50.	Cover All Course Requirements	.157
51.	Avoid Trivial Detail	.427**
52.	Answer Questions and Work Problems on Request	.352*
53.	Lecture over Students' Ability	.363*
54.	Give False Information about Course	.179
55.	Refuse to Explain Course Material	.247
56.	Follow Course Schedule	.125
57.	Prepare for Class	.363*
58.	Base Tests on Relevant Material	.359*
59.	Base Tests on Emphasized Material	.431**
60.	Make Tests Too Difficult	.363*
61.	Schedule Quizzes at Regular Intervals	.363*
62.	Make Tests Too Long	.215
63.	Comment on Returned Papers	.352*
64.	Excuse High Average Students from Final	.352*
65.	Permit Extra-Credit Work	.607**

Table 2 (continued)

66.	Disregard Lowest Test Score	.398*
67.	Refuse to Explain Grading System	.352*
68.	Tell How Students Are Graded	.262
69.	Curve Grades	.310
70.	Return All Papers	.363*
71.	Grade All Assignments	.363*
72.	Give Make-Up Tests at Students' Convenience	.363*
73.	Grade on Major, Sex, Athlete, etc.	.273
74.	Grade on Class Attendance	.352*
75.	Grade on Final Exam Only	.401*
76.	Pass/Fail a Predetermined Percentage	.172
77.	Willing to Discuss Grades	.295
78.	Derogate Teaching	.421**
79.	Derogate the Course	.352*
80.	Indicate Preference for Consulting or Research	.363*
81.	Criticize Fellow Teachers	.363*

$n = 38$.

*p 0.05.

**p 0.01.

showed some stability from class to class over professors, from a realistic standpoint the stability was not remarkably high for any behavior. More research is required to clarify this situation.

Validity

The only other study which employed a comparable method of validation was that of Douglass (1968). However, Douglass was less critical of the critical incident technique of gathering data than was this author. The checklist used by Douglass was gathered in much the same way as the checklist presented here (Appendix A). Three forms were compiled, one with an overall rating of teacher effectiveness added, one with an overall rating of teacher ability related to students' learning added, and one with both added. Items were placed into either a general effective category or a learning category. They were further judged as denoting effectiveness or ineffectiveness on the basis of the critical incident survey.

The three forms were administered by Douglass to seven classes each. Approximately 195 copies of each form were distributed. The checklist was scored by the ratio of effective to ineffective items; this overall scoring was deemed inappropriate for the present research since it depended so directly on the results of the critical incident survey which it was designed to corroborate. The learning criterion was the grade on a standardized final examination; as a control, student

grade point average was checked to be equal across classes and confirmed to be so. The results indicated that effectiveness rates were related to all criterion measures.

The present research considered only the relationship of individual items to the criterion of average grade in the class as adjusted for SAT v', SAT m', and HSA in an analysis of covariance. That is, this study attempted strictly to corroborate the results of the critical incident survey conducted by Ronan (1971). Of the 81 items, 3 were significantly related to the criterion at the 0.05 level (see Appendix B).

These three behavioral items were the following: (1) gives excessive work, (2) avoids trivial detail, and (3) talks too rapidly. Instructors who gave excessive work were judged by the criterion of teacher performance to be poorer teachers than those who did not give excessive work. Instructors who did not avoid trivial detail were judged to be poorer than those who did, and instructors who talked too rapidly were judged to be poorer than those who did not.

If this result were a true one, it is now necessary to determine whether these behaviors indeed affect a class differentially with regard to the previously mentioned teacher performance criterion. In particular, these three behaviors should be experimentally varied to assess their effects in the college classroom. Such systematic application is in order to determine whether or not the present results were due to chance factors, faulty design, or true causal relationships.

That 78 of the professorial behavior items on the checklist were not significantly related to the teacher performance criterion was indicative of the almost total failure of the critical incident technique as employed by Ronan (1971). Indeed, if these results were not attributable to occurrences such as chance errors accruing over statistical tests performed in the research, difficulties in the data collection, or a faulty design, it must be concluded that none of the 78 critical behaviors compiled by Ronan had differential effects upon the teacher performance criterion as would be expected if they were indeed critical.

That these previously discussed results were indeed true must be skeptically regarded for three major reasons. First, seven professors who taught classes eligible to be in Sample I declined to participate. Thus, Sample I was very probably a sample biased with respect to the effectiveness of the teachers contained in it.

Second, the Georgia Institute of Technology was itself biased with respect to its selection of professors and of students. The Georgia Institute of Technology selects very highly qualified professors and students, and thus, very little variance attributable to professorial behaviors might exist in such samples of professors as Sample I.

Finally, the experimental design may not have been appropriate for assessing validity. Light and Smith (1960) and Siegel and Siegel (1967) have noted that multivariate designs in which interactions may be assessed are indicated in assessing teacher performance. The as-

sumption was made in this research that all 81 behaviors were independent, and this very possibly restricted the results of the research. It must be noted, however, that Sample I was so small as to make assessing interactions of the various behaviors impossible. The experimental design also did not take class size into account although Holland (1954) demonstrated it to be an important variable in educational research. However, the effect of class size was impossible to assess because all classes in Sample I had approximately the same number of students.

Certainly, systematic replication would serve to clarify all the aforementioned results and is a very much needed step.

CHAPTER V

SUMMARY

The key to the rationale of this study was the belief that specific teacher behaviors occur in the classroom situation, that these teacher behaviors can be determined reliably, and that their relationship to a criterion of effectiveness can yield valuable information. As Krasner (1971) and Mischel (1968) both stressed in assessing performance, be it for clinical purposes or otherwise, it is the actual behavior in the situation in which it naturally occurs that is of interest. The critical incidents as gathered and compiled by Ronan (1971) and used in this research represent, as nearly as can be determined, a taxonomy of college classroom teacher behavior. This research was concerned with assessing these behaviors with respect to several psychometric considerations and with attempting to place teacher evaluation on a sound, methodological base as discussed by Arden (1968). A special effort was made in the first chapter to elucidate findings relating to the usefulness of ratings. It was concluded that ratings were untenable criteria of performance and were useless in evaluating teachers. Thus, despite the past history of the overwhelming use of ratings as teacher evaluation criteria, they were avoided in this research.

A major finding of this study was that students in a given class can agree to a very high degree on the behaviors that their professor exhibited in that class as represented by the present checklist (see Appendix A). This checklist, which was not a rating as such, thus exhibited high perceptual stability, a nonstandard form of inter-observer reliability original with this research. This high perceptual stability was in stark contrast to results using ratings, items which ask an observer to make a subjective judgment in relation to a vaguely defined continuum. Ryans (1954) has noted that teachers were quite aware of the shortcoming of ratings with regard to inter-observer reliability.

Next, it was found that most behaviors of professors as represented by items on the checklist exhibited a statistically significant ($p = 0.05$) degree of consistency from class to class although this consistency was not impressive in any given case. This finding might be taken to indicate that assessments of a teacher's performance may not be feasible or simple except within a given class.

Finally, it was found that only 3 of the 81 teacher behaviors assessed by the present checklist were significantly ($p = 0.05$) related to the teacher performance criterion employed in this study. The criterion for a given professor was the average grade in his class adjusted for covariation due to aptitude and achievement levels of the students. This possibly indicates that college teachers' behaviors influence their

students' gains in knowledge only slightly. However, there were probably biases in the sample of professors and of students which influenced the results; careful replication is in order to determine whether this was indeed the case.

APPENDIX A

RONAN'S (1971) CHECKLIST

GEORGIA INSTITUTE OF TECHNOLOGY
EXPERIMENTAL TEACHER IMPROVEMENT QUESTIONNAIRE

This form is to be returned for later re-use. Record your answers on the accompanying computer answer sheet. Add any comments on the reverse of this answer sheet, away from the area used for the answers. Do not sign your name.

Thank you for your assistance.

TEACHER EVALUATION On answer sheet, T = yes, F = No

1. Know or attempt to know students' names?
2. Talk with students before and/or after class?
3. Hold social events for his students?
4. Refuse to give advice or assistance on student request (class or office) with personal problems?
5. Discuss (answer questions on) extra-class issues?

6. Encourage (answer) all questions in class?
7. Treat all students equally (regardless of sex, major, etc.)?
8. Ridicule, "ride," or otherwise embarrass students (either on questions or their performance)?
9. Give or offer individual help with course material (class or office)?
10. Lose control of himself in class (shout, curse, anger, etc.)?

11. Bother (harass) students during recitation, quizzes, etc.?
12. Make threats concerning class work or personal behavior?
13. Accept your excuse, explanation (as for missing quiz)?
14. Refuse to listen to or recognize other viewpoints in class?
15. Say or indicate in some way that students are inferior?

16. Miss more than two scheduled (re-scheduled) classes?
17. On time for all classes?
18. Arrange quiz dates or dead-lines for student convenience?
19. End lectures at end of class time?
20. Distribute a course outline or study plan (course objectives)?

21. Follow course outline or study plan?
22. Give examples of quiz items?
23. Require and grade homework?

24. Return papers and quizzes promptly?
25. Permit classroom disturbances (as students talking to each other)?

26. Make false statements concerning course requirements (number of cuts, grading, etc.)?
27. Give excessive work?
28. Ask student preference as to topics covered?
29. Ask for student suggestions on his teaching?
30. Encourage (ask for) discussion, questions, or student opinions?

31. Appear well-groomed?
32. Speak clearly and distinctly?
33. Use humor in lecture to illustrate points?
34. Read lectures from notes or book?
35. Appear nervous, ill-at-ease during lecture?

36. Talk or present material too rapidly?
37. Give rambling, dis-organized lecture?
38. Look at students during lecture?
39. Use language students understand?
40. Use profane language excessively?

41. Stress, in some way, important points in the material?
42. Use current, pertinent, or personal examples to illustrate points?
43. Show usefulness of material in "real world"?
44. Admit not knowing answer to a question?
45. Use outside references to supplement course?

46. Distribute hand-outs/notes to supplement lecture?
47. Use visual aids to (including blackboard) supplement lecture?
48. Have full command of the subject matter?
49. Usually lectures from text?
50. Cover all course requirements?

51. Avoid trivial detail?
52. Answer questions; work problems if requested?
53. Lecture over students' heads?
54. Give erroneous information about course material?
55. Refuses to or does not explain course material?

56. Follow course schedule?
57. Prepared for class?
58. Base tests on relevant (important) material?
59. Base test on emphasized material?
60. Make tests too difficult?

61. Schedule quizzes at regular intervals?
62. Make tests too long?
63. Comment on (correct) returned papers, quizzes, etc.?
64. Excuse high average students from final?
65. Permit extra work to improve grade?

66. Disregard lowest test score in grading?
67. Refuse to explain grading system?
68. Tell how students are to be graded?
69. Curve grades?
70. Return all papers and quizzes?

71. Grade all quizzes and assignments?
72. Give make-up tests at individual convenience?
73. Grade on such things as major, sex, athlete, etc.?
74. Grade on class attendance?
75. Grade on final exam only?

76. Pass/fail a predetermined percentage of the class?
77. Willing to discuss grades?
78. Make derogatory comments about teaching?
79. Make derogatory comments about the course?
80. Indicate he would rather consult and/or do research than teach?

81. Make personal criticisms of fellow teachers?

APPENDIX B

ANALYSES OF COVARIANCE OF THE 81 ITEMS

Table 3. Item 1: Know Students' Names

Source	SS	DF	MS	F	P less than
Within Cells	670.748	8	83.844		
Regression	909.765	3	303.255	3.617	.065
F	185.009	1	185.009	2.207	.176

Table 4. Item 5: Discuss Extraclass Issues

Source	SS	DF	MS	F	P less than
Within Cells	760.935	8	95.117		
Regression	1018.114	3	339.371	3.568	.067
F	92.630	1	92.630	.974	.353

Table 5. Item 6: Encourage All Questions

Source	SS	DF	MS	F	P less than
Within Cells	770.770	8	96.346		
Regression	990.680	3	330.227	3.427	.073
F	84.987	1	84.987	.882	.375

Table 6. Item 17: On Time for Class

Source	SS	DF	MS	F	P less than
Within Cells	727.117	8	90.890		
Regression	979.016	3	326.339	3.590	.066
F	128.641	1	128.641	1.415	.268

Table 7. Item 18: Arrange Deadlines for Student Convenience
Item 61: Schedule Quizzes at Regular Intervals

Source	SS	DF	MS	F	P less than
Within Cells	848.121	8	106.015		
Regression	949.042	3	316.347	2.984	.096
F	7.636	1	7.636	.072	.795

Table 8. Item 19: End Class on Time

Source	SS	DF	MS	F	P less than
Within Cells	852.314	8	106.539		
Regression	771.950	3	257.317	2.415	.142
F	3.444	1	3.444	.032	.862

Table 9. Item 20: Distribute Course Outline

Source	SS	DF	MS	F	P less than
Within Cells	848.047	8	106.006		
Regression	855.983	3	258.325	2.692	.117
F	5.518	1	5.518	.052	.825

Table 10. Item 21: Follow Course Outline

Source	SS	DF	MS	F	P less than
Within Cells	646.839	8	80.855		
Regression	983.359	3	327.786	4.054	.050
F	208.919	1	208.919	2.584	.147

Table 11. Item 22: Give Examples of Quiz Items

Source	SS	DF	MS	F	P less than
Within Cells	837.730	8	104.716		
Regression	961.916	3	320.639	3.062	.091
F	18.027	1	18.027	.172	.689

Table 12. Item 23: Require and Grade Homework

Source	SS	DF	MS	F	P less than
Within Cells	827.540	8	103.442		
Regression	808.167	3	269.389	2.604	.124
F	28.218	1	28.218	.273	.616

Table 13. Item 24: Return Papers and Quizzes Promptly

Source	SS	DF	MS	F	P less than
Within Cells	787.575	8	98.447		
Regression	576.924	3	192.308	1.953	.200
F	68.182	1	68.182	.693	.429

Table 14. Item 25: Permit Classroom Disturbances

Source	SS	DF	MS	F	P less than
Within Cells	804.587	8	100.573		
Regression	950.520	3	316.840	3.150	.086
F	48.978	1	48.978	.487	.505

Table 15. Item 27: Give Excessive Work
Item 51: Avoid Trivial Detail

Source	SS	DF	MS	F	P less than
Within Cells	484.234	8	60.529		
Regression	962.730	3	320.910	5.302	.026
F	371.523	1	371.523	6.138	.038

Table 16. Item 30: Encourage Discussions

Source	SS	DF	MS	F	P less than
Within Cells	853.446	8	106.681		
Regression	947.853	3	315.951	2.962	.098
F	.119	1	.119	.001	.974

Table 17. Item 32: Speak Distinctly

Source	SS	DF	MS	F	P less than
Within Cells	853.604	8	106.700		
Regression	611.268	3	203.756	1.910	.207
F	2.154	1	2.154	.020	.891

Table 18. Item 33: Use Humor

Source	SS	DF	MS	F	P less than
Within Cells	815.793	8	101.937		
Regression	929.682	3	309.894	3.040	.093
F	40.264	1	40.264	.395	.547

Table 19. Item 34: Read Lectures from Notes

Source	SS	DF	MS	F	P less than
Within Cells	853.604	8	106.700		
Regression	611.268	3	203.756	1.910	.207
F	2.154	1	2.154	.020	.891

Table 20. Item 36: Talk Too Rapidly

Source	SS	DF	MS	F	P less than
Within Cells	433.253	8	54.157		
Regression	1369.811	3	456.604	8.431	.007
F	422.505	1	422.505	7.802	.023

Table 21. Item 42: Use Current, Pertinent Examples

Source	SS	DF	MS	F	P less than
Within Cells	700.851	8	87.606		
Regression	902.711	3	300.904	3.435	.072
F	154.907	1	154.907	1.768	.220

Table 22. Item 43: Show Usefulness of Material

Source	SS	DF	MS	F	P less than
Within Cells	849.802	8	106.225		
Regression	933.588	3	311.196	2.930	.100
F	5.956	1	5.956	.056	.819

Table 23. Item 45: Use Outside References

Source	SS	DF	MS	F	P less than
Within Cells	840.823	8	105.103		
Regression	947.437	3	315.812	3.005	.095
F	12.742	1	12.742	.121	.737

Table 24. Item 46: Distribute Lecture Supplements

Source	SS	DF	MS	F	P less than
Within Cells	629.932	8	78.741		
Regression	1034.250	3	344.750	4.378	.042
F	225.826	1	225.826	2.868	.129

Table 25. Item 49: Lecture from Test

Source	SS	DF	MS	F	P less than
Within Cells	832.978	8	104.122		
Regression	957.457	3	319.152	3.065	.091
F	22.779	1	22.779	.219	.652

Table 26. Item 53: Lecture over Students' Heads

Source	SS	DF	MS	F	P less than
Within Cells	669.047	8	83.631		
Regression	1100.761	3	366.920	4.387	.042
F	186.711	1	186.711	2.233	.173

Table 27. Item 58: Base Tests on Relevant Material

Source	SS	DF	MS	F	P less than
Within Cells	713.051	8	89.131		
Regression	1041.650	3	347.217	3.896	.055
F	142.706	1	142.706	1.601	.241

Table 28. Item 59: Base Tests on Emphasized Material

Source	SS	DF	MS	F	P less than
Within Cells	821.230	8	102.654		
Regression	824.502	3	274.834	2.677	.118
F	34.528	1	34.528	.336	.578

Table 29. Item 60: Make Tests Too Difficult

Source	SS	DF	MS	F	P less than
Within Cells	840.734	8	105.092		
Regression	951.385	3	317.128	3.018	.094
F	12.832	1	12.832	.122	.736

Table 30. Item 62: Make Tests Too Long

Source	SS	DF	MS	F	P less than
Within Cells	827.978	8	103.497		
Regression	977.661	3	325.887	3.149	.086
F	27.780	1	27.780	.268	.618

Table 31. Item 66: Disregard Lowest Test Score

Source	SS	DF	MS	F	P less than
Within Cells	828.708	8	103.588		
Regression	842.091	3	280.697	2.710	.115
F	27.050	1	27.050	.261	.623

Table 32. Item 68: Tell How Students Are Graded

Source	SS	DF	MS	F	P less than
Within Cells	849.759	8	106.220		
Regression	589.686	3	196.562	1.851	.216
F	5.999	1	5.999	.056	.818

Table 33. Item 69: Curve Grades

Source	SS	DF	MS	F	P less than
Within Cells	752.346	8	94.043		
Regression	806.668	3	268.889	2.859	.104
F	101.219	1	101.219	1.076	.330

Table 34. Item 74: Grade on Attendance

Source	SS	DF	MS	F	P less than
Within Cells	752.704	8	94.088		
Regression	1008.746	3	336.249	3.574	.066
F	100.861	1	100.861	1.072	.331

Table 35. Items Which Exhibited No Interclass Differences

Item No.	Behavior
2	Talk with students outside of class.
3	Hold social events.
4	Refuse to advise on personal problems.
7	Treat all students equally.
8	Ridicule students.
9	Offer individual help.
10	Lose temper.
11	Harass students.
12	Make threats.
13	Accept excuses for missing a quiz.
14	Refuse to recognize other viewpoints.
15	Indicate students are inferior.
16	Miss more than two classes.
26	Make false statements concerning course requirements.
28	Ask students preference for topics covered.
29	Ask for student suggestions concerning teaching.
31	Appear well-groomed.
35	Appear nervous.
37	Give rambling lectures.
38	Look at students while lecturing.
39	Use language students understand.
40	Use profane language.
41	Stress important points.
44	Admit not knowing answer.
47	Use visual aids.
48	Know subject matter.
50	Cover course requirements.
52	Answer questions; work problems.
54	Give false information about course material.
55	Refuse to explain course material.
56	Follow course schedule.
57	Prepared for class.
63	Comment on returned papers.
64	Excuse high average students from final
65	Permit extra-credit work.
67	Refuse to explain grading system.
70	Return all papers.
71	Grade all assignments.
72	Give make-up tests at individual convenience.
73	Grade on sex, major, athlete, etc.
75	Grade on final only.

Table 35 (continued)

76	Pass/fail a predetermined percentage.
77	Willing to discuss grades.
78	Derogate teaching.
79	Derogate the course.
80	Indicate he prefers consulting or research.
81	Criticize fellow teachers.

BIBLIOGRAPHY

Aiken, L. R. The grading behavior of college faculty. Educational and Psychological Measurement, 1963, 23, 319-322.

Arden, E. Faculty as teachers. Educational Forum, 1968, 32, 447-452.

Astin, A. W. Undergraduate achievement and institutional "excellence." Science, 1968, 161, 661-668.

Baehr, M. E., & Williams, G. B. Prediction of success from factorially determined dimensions of personal background data. Journal of Applied Psychology, 1968, 52, 98-103.

Barr, A. S., Bechdolt, B. B., Coxe, W. W., Gage, N. L., Orleans, J. J., Hemmers, H. H., & Ryans, D. C. Report of the committee on the criteria of teacher effectiveness. Review of Educational Research, 1953, 22, 238-263.

Barr, A. S., & Jones, R. The measurement and prediction of teacher efficiency. Review of Educational Research, 1958, 27, 256-264.

Besco, R. O., & Lawshe, C. H. Foreman leadership as perceived by superiors and subordinates. Personnel Psychology, 1959, 12, 573-582.

Borg, W. R., & Hamilton, E. R. Comparison between a performance test and criteria of instructor effectiveness. Psychological Reports, 1956, 2, 111-116.

Braunhaussen, N. Selection des employes de bureau. Revue de Science du Travail, 1929, 1, 499-512.

Bruner, J. S. Social psychology and perception. In E. E. Maccody, T. R. Newcomb, and E. L. Hartley (Eds.), Readings in social psychology, New York: Holt, 1958.

Campbell, D. T., & Fiske, D. W. Convergent and discriminant validation by the multitrait multimethod matrix. Psychological Bulletin, 1959, 56, 81-105.

Charest, A. G., Cowart, D. G., & Goodman, P. S. Multi-instrument, multi-rater, multi-trait method for assessing measures of managerial performance. Experimental Publication System, 1963, 3, No. 092A.

Cook, W. W., & Leeds, C. H. Measuring the teaching personality. Educational and Psychological Measurement, 1947, 7, 399-410.

Crawford, P. L., & Bradshaw, H. L. Perception of characteristics of effective teachers: A scaling analysis. Educational and Psychological Measurement, 1968, 28, 1079-1085.

Cureton, E. E. Validity. In E. F. Lindquist (Ed.), Educational Measurement, Washington, D. C.: American Council on Education, 1951.

Deshpande, A. S., Webb, S. C., & Marks, E. Student perception of engineering instructor behaviors and their relationships to the evaluations of instructors and courses. American Educational Research Journal, 1970, 7, 289-305.

Domas, S. J., & Tiedeman, D. V. Teacher competence: An annotated bibliography. Journal of Experimental Education, 1950, 19, 101-218.

Douglass, L. C. Measures of teacher evaluation as related to student achievement. Unpublished masters thesis, University of Tennessee, 1968.

Eble, K. E. Special report: Project to improve college teaching. Washington, D. C.: American Association of University Professors (undated).

Eckert, R. E. Ways of evaluating college teaching. School and Society, 1950, 71, 65-69.

Eells, W. C. College teachers and college teaching: An annotated bibliography. Atlanta: Southern Education Board, 1957.

Elliot, D. N. Characteristics and relationships of various criteria of college and university teaching. West Lafayette, Ind.: Purdue University, 1950.

Flanagan, J. C. The aviation psychology program in the Army Air Force, report no. 1. Washington, D. C.: U. S. Government Printing Office, 1948.

Flanagan, J. C. Job requirements. In W. Dennis (Ed.), Current trends in industrial psychology. University of Pittsburg, 1949. (a)

Flanagan, J. C. A new approach to evaluating personnel. Personnel, 1949, 26, 35-42. (b)

Flanagan, J. C. The critical incident technique. Psychological Bulletin, 1954, 51, 327-358.

Fleishmen, E. A., Harris, E. F., & Burtt, H. E. Leadership and supervision in industry. Bureau of Educational Research Monograph, 1955, No. 33.

Goldberg, L. R. Research on the clinical judgment process. N. I. P. P. Paper read at Clinical Section meeting, Nymegen, Netherlands, 1967.

Graham, W. K., & Calendo, J. T. Personality correlates of supervising ratings. Personnel Psychology, 1969, 22, 483-487.

Gustad, J. W. Policy and practices in faculty evaluation. Educational Record, 1961, 42, 194.

Gustad, J. W. On improving college teaching. National Education Association Journal, 1964, 53, 37.

Heilman, J. D., & Armentrout, W. D. The rating of college teachers on ten traits by their students. Journal of Educational Psychology, 1936, 27, 197.

Hills, J. R. The effect of admissions practices on college grading standards. Journal of Educational Measurement, 1964, 2, 115-118.

Holland, J. B. The image of the instructor as it is related to class size. Journal of Experimental Education, 1954, 23, 171-177.

Jones, E. E., & deCharms, D. Changes in social perception as a function of the personal relevance of behavior. Human Relations, 1957, 20, 75-85.

Jones, E. E., & Thibaut, J. W. The organizing function of interaction roles in person perception. Journal of Abnormal and Social Psychology, 1958, 57, 155-164.

Kavanagh, M. J., MacKinney, A. C., & Wolins, L. Issues in managerial performance: Multitrait-multimethod analysis of ratings. Psychological Bulletin, 1971, 75, 34-39.

Kipnis, D. Some determinants of supervisory esteem. Personnel Psychology, 1960, 13, 377-391.

Kirchner, W. K. Relationships between supervisory and subordinate ratings for technical personnel. Journal of Industrial Psychology, 1966, 3, 57-60.

Konigsburg, D. Development and preliminary evaluation of an instructor checklist based on the critical technique. Pittsburgh: Doctoral Dissertation Series, 1954, #7937.

Krasner, L. Behavior therapy. In P. H. Mussen, and M. R. Rosenzweig (Eds.), Annual Review of Psychology. Palo Alto, California: Annual Reviews, Inc., 1971.

Latham, G. P. The development of job performance criteria for pulpwood producers in the southeastern United States. Atlanta: Unpublished masters thesis, Georgia Institute of Technology, 1969.

Lifson, K. A. Errors in time-study judgments of industrial work pace. Psychological Monographs: General and Applied, 1953, 67, No. 355.

Light, R. L., & Smith, P. V. Choosing a future: Strategies for designing and evaluating new programs. Harvard Educational Review, 1960, 40, 1-28.

Locke, E. A. The development of criteria of student achievement. Educational and Psychological Measurement, 1963, 2, 299-308.

Mager, R. F. Preparing instructional objectives. Belmont, California: Fearon, 1962.

McGrath, E. J. Characteristics of Outstanding College Teachers. Journal of Higher Education, 1962, 33, 148.

Mischel, W. Personality and assessment. New York: John Wiley & Sons, 1968.

Prien, E. P. Development of a supervisor position description questionnaire. Journal of Applied Psychology, 1962, 47, 10-14.

Prien, E. P. Measuring performance of bank tellers. Experimental Publication System, 1969, 3, No. 095C.

Prien, E. P., & Liske, R. E. Assessment of higher level personnel: III Rating criteria: A comparative analysis of supervisory ratings and incumbent self-rating of job performance. Personnel Psychology, 1962, 15, 187-194.

Remmers, J. J. Rating methods in research on teaching. In N. L. Gage (Ed.), Handbook of research on teaching. Chicago: Rand McNally, 1963.

Ronan, W. W. A factor analysis of eight job performance measures. Journal of Industrial Psychology, 1963, 1, 107-112. (a)

Ronan, W. W. A factor analysis of eleven job performance measures. Personnel Psychology, 1963, 16, 255-267. (b)

Ronan, W. W. Development of an instrument to evaluate college classroom teaching effectiveness. Atlanta: Georgia Institute of Technology, Department of Health, Education, and Welfare Grant No. 1-2-045, 1971.

Ronan, W. W., & Prien, E. P. Toward a criterion theory: A review and analysis of research and opinion. Greensboro, N. C.: The Richardson Foundation, 1966.

Ronan, W. W., & Prien, E. P. Perspectives on the measurement of human performance. New York: Appleton-Century-Crofts, 1971.

Rosenthal, R. Experimenter effects in behavioral research. New York: Appleton-Century-Crofts, 1966.

Ryans, D. G. Notes on the rating of teacher performance. Journal of Educational Research, 1954, 47, 695-703.

Safren, M., & Chapanis, A. A critical incident study of hospital medication errors. Hospitals, 1960, 34, 32-34.

Schneider, B., & Bartlet, C. J. Industrial differences and organizational climate: II Measurements of the organizational climate by the multi-trait, multi-rater matrix. Personnel Psychology, 1970, 23, 493-512.

Siegel, A. I. The checklist as a criterion of proficiency. Journal of Applied Psychology, 1954, 38, 93-95.

Siegel, L., & Siegel, L. C. A multivariate paradigm for behavior research. Psychological Bulletin, 1967, 68, 306-326.

Smit, J. A. A study of the critical requirements for instructors of general psychology courses. Pittsburgh: University of Pittsburgh, 1952, 48, 279-284.

Springer, D. Ratings of candidates for promotion by coworkers and supervisors. Journal of Applied Psychology, 1953, 37, 347-351.

Strander, N. E. A longitudinal study of some relationships among criteria of managerial performance as perceived by supervisors and subordinates. Journal of Industrial Psychology, 1965, 3, 43-51.

Taft, R. The ability to judge people. Psychological Bulletin, 1955, 52, 1-23.

Thorndike, R. L. Personnel selection. New York: John Wiley & Sons, 1949.

Tiffin, J. T., & McCormick, E. J. Industrial Psychology (5th ed.). Englewood Cliffs, New Jersey: Prentice-Hall, 1965.

Webb, S. C. Increased selectivity and institutional standards. In K. M. Wilson (Ed.), Research related to college admissions, Atlanta: Southern Regional Board, 1963.

Webb, S. C. Student perceptions of instructor teaching goods in correspondence with instructor self-ratings. Atlanta: Georgia Institute of Technology, Office of Evaluation Studies, 1967.